Regresión y correlación lineal

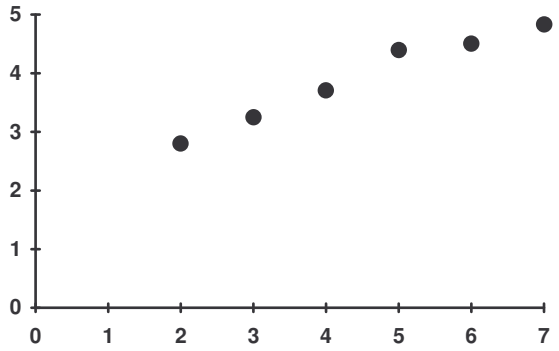
Curvas de regresión

Una vez que hemos establecido la existencia de una relación estadística entre dos variables y determinado la intensidad de esta relación, el siguiente paso es ver como pueden predecirse los valores de una variable en función de los de otra y que grado de precisión tendrán estas predicciones. A estas cuestiones atiende el término regresión, introducido por Sir Francis Galton al estudiar la relación entre la estatura de padres e hijos.

Se llama curva de regresión de una variable estadística Y sobre otra variable X, a la curva que se obtiene representando las medias condicionadas $\bar{y}_i$ en función de los valores x_i de la variable X. Se tratará de una verdadera curva, si la variable X es continua, o de una sucesión de puntos si la variable es discreta. Por ejemplo, volviendo a los datos de la tarea de Sternberg, utilizados en el capítulo anterior, teníamos que las medias del tiempo de respuesta para las distintas longitudes de la lista eran:

$$\bar{y}_2 = 2,8 \quad \bar{y}_3 = 3,25 \quad \bar{y}_4 = 3,71 \quad \bar{y}_5 = 4,4 \quad \bar{y}_6 = 4,5 \quad \bar{y}_7 = 4,83$$

con lo cual al hacer la representación gráfica tendríamos:



La curva de regresión de Y sobre X visualiza como cambia la media de la variable Y de aquellos grupos de observaciones caracterizados por tener un mismo valor en la otra variable X. Es decir, como varía, por término medio, la variable Y en función de los valores de X. Por ello la variable Y recibe el nombre de variable dependiente y la variable X el de variable independiente. En nuestro ejemplo sería, como cambia, en promedio, el tiempo de respuesta en función de la longitud de la lista.

De manera análoga, se llama curva de regresión de X sobre Y a la curva obtenida representando las medias condicionadas $\bar{x}_j$ en función de los valores y_j. En este caso la variable dependiente sería X y la independiente Y.

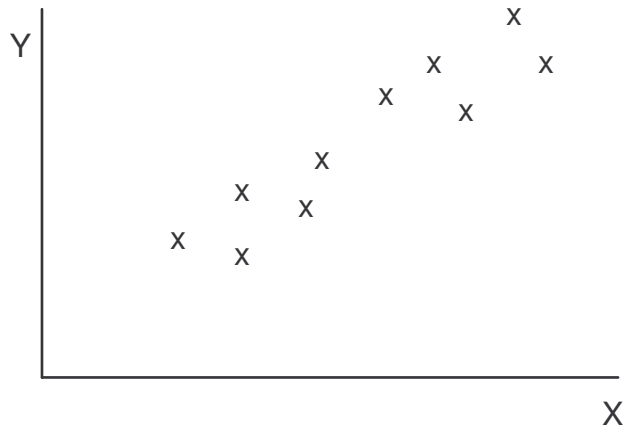
Hemos dicho que la curva de regresión será una verdadera curva, cuando la variable independiente sea continua. Pero construirla requeriría tener información sobre todos sus posibles valores, que al ser infinitos, lo hacen imposible. En realidad lo que tendremos será información acerca de los valores que toma la variable dependiente, para un conjunto finito de valores de la variable independiente. Es decir, tendremos un

conjunto, más o menos numeroso, pero finito, de puntos del plano a partir del cual debemos calcular o aproximar la curva de regresión para cualquier valor de la variable independiente.

Para resolver esta cuestión, recurriremos al procedimiento de ajuste.En él se fija, por criterios externos, la forma de la función que relaciona ambas variables: una recta, una función cuadrática, etc. y se determina la función concreta que mejor se ajusta a los datos existentes, considerándose las discrepancias entre los valores que proporciona esta función para la variable dependiente, y los obtenidos, como errores o residuos debidos a la fluctuación muestral, o a factores incontrolados. Se entenderá que la función que mejor se ajusta, será aquella que haga mínimos, en algún sentido, los errores mencionados.

Regresión lineal

En la regresión lineal partimos de un conjunto de observaciones (x_i y_i) de dos variables X e Y. Por alguna razón, hemos decidido que la relación existente entre ambas variables es esencialmente lineal. Las razones para esta decisión pueden estar basadas en la inspección de la representación gráfica de los datos, en consideraciones teóricas sobre la forma en que actúa una variable sobre otra, ser una primera aproximación tentativa, o una mezcla de todas las anteriores. Aunque la relación entra esas dos variables sea substancialmente lineal, las observaciones que obtenemos no estarán perfectamente alineadas, sino que se encontrarán más o menos dispersas debido a factores no controlados o a la variabilidad intrínseca de la variable considerada. La situación será la que se refleja en la figura siguiente:



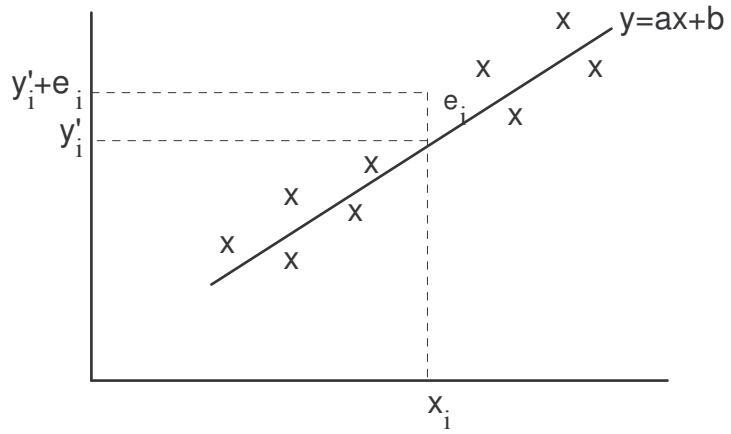
Si pretendemos predecir la variable Y en función de los valores de la variable X deberemos de determinar la recta de regresión de Y sobre X que tendrá como expresión genérica:

$$y = a\,x + b$$

donde a es la pendiente de la recta que representa el incremento que experimenta el valor y por cada unidad que se incrementa el valor de x, y b es la ordenada en el origen, es decir el valor de y cuando x vale cero.

Obviamente, como los puntos que representan a las observaciones, no están perfectamente alineados, la recta no determinará exactamente los valores de y. Si

calculásemos la diferencia entre el valor observado y el que predice la recta, obtendríamos una diferencia que llamaremos residuo. Esta situación puede verse en el gráfico siguiente:



Si designamos y' los valores que predice la recta de regresión:

$$y_i' = ax_i + b$$

de acuerdo con lo expresado anteriormente, los residuos serán:

$$e_i = y_i - y_i'$$

y despejando obtenemos el modelo explicativo de las observaciones:

$$y_i = y_i' + e_i = ax_i + b + e_i$$

Recta de mínimos cuadrados

Para abordar el problema de determinar los coeficientes a y b de la recta que mejor se ajusta a un conjunto de observaciones (x_i y_i), utilizaremos el método de mínimos cuadrados. Este método consiste en determinar la recta de tal forma que la suma de los cuadrados de los residuos sea mínima. Aplicando este método se obtienen, como puede verse en el apéndice matemático, las siguientes expresiones:

$$a = \frac{S_{XY}}{S_X^2}$$

$$b = \bar{y} - a\bar{x}$$

Ejemplo: Volviendo a los datos de la muestra de 15 estudiantes, a los cuales se había pasado un test de inteligencia, cuyas puntuaciones se reflejan en la variable X, y a los que se había realizado una prueba que se reflejaba en las puntuaciones de la variable Y, vamos a calcular la recta de regresión de Y sobre X. Es decir, la recta que permite predecir las puntuaciones en la prueba, a partir de los valores del test de inteligencia. La tabla siguiente nos proporciona los valores de estas variables, así como los cálculos básicos necesarios para determinar los coeficientes de la recta:

Sujeto	x_i	y_i	$x_i \cdot y_i$	x_i^2
1	9	5	45	81
2	12	5	60	144
3	6	1	6	36
4	9	4	36	81
5	7	2	14	49
6	9	2	18	81
7	5	1	5	25
8	9	3	27	81
9	7	3	21	49
10	3	1	3	9
11	10	4	40	100
12	6	2	12	36
13	11	5	55	121
14	4	2	8	16
15	13	5	65	169
Totales	120	45	415	1078

La media de X es 120/15 = 8
La media de Y es 45/15 = 3
La covarianza es 415/15 - 24 = 3,67
La varianza de X es 1078/15 - 64 = 7,87

Por consiguiente, la pendiente de la recta será:

$$a = 3{,}67/7{,}87 = 0{,}466$$

y la ordenada en el origen:

$$b = 3 - 0{,}466 \cdot 8 = 3 - 3{,}728 = -0{,}728$$

por consiguiente la expresión de la recta de regresión de Y sobre X será:

$$y = 0{,}466x - 0{,}728$$

Predicción, varianza explicada

Como ya dijimos, el objetivo de la regresión es poder predecir los valores de una variable en función de los de la otra variable. De esta forma los valores que proporciona la recta de regresión de Y sobre X:

$$y_i' = ax_i + b$$

deben de entenderse como las predicciones de la variable Y que se obtienen para cada valor x_i. Esto permite obtener valores de dicha variable, incluso en aquellos casos para los cuales no existen observaciones. Por ejemplo, en los datos del test de inteligencia, anteriormente usado, no existía ningún sujeto con ocho puntos en el test, a pesar de ello podemos predecir que puntuación le correspondería en la prueba de rendimiento a esa puntuación, sustituyendo en la recta de regresión:

$$y' = 0{,}466\, x - 0{,}728$$

x por el valor 8, tenemos:

$$y' = 0{,}466 \cdot 8 - 0{,}728 = 3$$

esta predicción debe entenderse, no como el valor que alcanzaría en la prueba de rendimiento un sujeto que tenga una puntuación de 8, sino como la media de los valores de todos los sujetos con puntuación de 8. Téngase en cuenta que el modelo explicativo de las observaciones:

$$y_i = ax_i + b + e_i$$

incluye un término individual y aleatorio, el residuo e_i, que proporciona la discrepancia entre el valor individual y la predicción común a todos individuos con el mismo valor x_i. Por ejemplo, en los datos tenemos cuatro individuos con 9 puntos en el test de inteligencia, el 1º, 4º, 6º y 8º, para todos ellos la predicción de la recta de regresión sería la misma:

$$y' = 0{,}466 \cdot 9 - 0{,}728 = 3{,}466$$

sin embargo sus valores empíricos son $y_1 = 5$, $y_4 = 4$, $y_6 = 2$, $y_8 = 3$, con lo cual los residuos correspondientes serían: $e_1 = 5 - 3{,}466 = 1{,}534$; $e_4 = 4 - 3{,}466 = 0{,}534$; $e_6 = 2 - 3{,}466 = -1{,}466$ y $e_8 = 3 - 3{,}466 = -0{,}466$.

Como los residuos son la diferencia entre los valores empíricos y_i y la predicción y' que corresponde a un valor x_i, cabe esperar que su media sea cero, por consiguiente la media de los valores y_i correspondientes a los individuos con valor x_i, será igual a la predicción común y'_i, como esto sucede para todo valor x_i, entonces la media de la variable Y coincidirá con la media de las predicciones.

A partir de la expresión:

$$y_i = y'_i + e_i$$

y considerando que los valores de los residuos son independientes de los valores x_i y por tanto las predicciones serán independientes de los residuos, puede deducirse que la varianza de la variable Y se descompone en la suma de la varianza de las predicciones más la varianza de los residuos, es decir:

$$S_Y^2 = S_{y'}^2 + S_e^2$$

La varianza de las predicciones es la parte de la variabilidad de la variable Y que puede determinarse a partir de la relación con la variable X. Mientras que la varianza de los residuos es la parte de la variabilidad de la variable Y, que no puede ser explicada por su relación con la variable X. Por tanto, la proporción de la varianza explicada:

$$\frac{S_{y'}^2}{S_Y^2}$$

será un indicador del grado de ajuste de los datos a la recta de regresión obtenida y de la exactitud de las predicciones que se realicen, a partir de ella.

Varianza explicada, coeficiente de determinación

Si tenemos en cuenta que la expresión que nos proporciona las predicciones es:

$$y'_i = ax_i + b$$

vemos que de hecho las predicciones son una transformación lineal de los valores de la variable independiente X, es decir que cada predicción se obtiene multiplicando el correspondiente valor x_i por una constante a y sumándole otra constante b. Si recordamos las propiedades de la varianza, los cambios de origen, sumar a todos los valores de la variable una misma cantidad, no afectaban a la varianza. Mientras que los cambios de escala, multiplicar todos los valores de la variable por una mismo número, hacían que la varianza quedase multiplicada por el cuadrado del factor de escala. Por consiguiente, la varianza de las predicciones será igual a la varianza de la variable independiente por el cuadrado de la pendiente. Es decir:

$$S^2_{y'} = a^2 S^2_X$$

si en la anterior igualdad sustituimos a por la expresión que permite su cálculo, tendremos:

$$a = \frac{S_{XY}}{S^2_X} \Rightarrow S^2_{y'} = \frac{S^2_{XY}}{S^4_X} \cdot S^2_X = \frac{S^2_{XY}}{S^2_X}$$

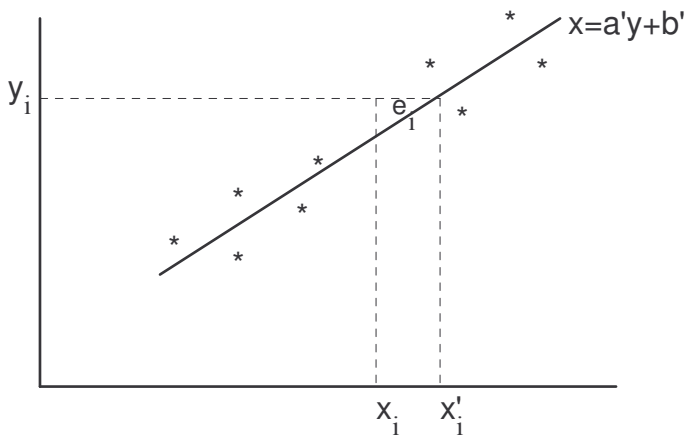
para calcular la proporción de la varianza explicada, dividimos por la varianza total de la variable dependiente Y, obteniendo:

$$\frac{S^2_{y'}}{S^2_Y} = \frac{S^2_{XY}}{S^2_X S^2_Y} = r^2$$

luego, la proporción de la varianza explicada es igual al cuadrado del coeficiente de correlación de Pearson, cantidad que recibe el nombre de coeficiente de determinación. Y que obviamente se interpreta como la proporción de la varianza de la variable dependiente Y que se explica por su relación lineal con la variable independiente X.

Recta de regresión de X sobre Y

Hasta ahora todo lo que hemos descrito hacía referencia a la predicción de la variable Y a partir de los valores de la variable X, pero que ocurriría si deseásemos predecir la variable X en función de los valores de la variable Y. Aunque los datos de los que disponemos son los mismos, la recta de regresión de X sobre Y va a ser diferente en función de que los residuos que se van a minimizar son distintos, gráficamente se aprecia fácilmente esta diferencia:



en este caso los residuos, cuya suma de cuadrados ha de minimizarse, son horizontales pues se refieren a discrepancias en el valor de x, mientras que en el caso de la recta de regresión de Y sobre X eran verticales pues correspondían a diferencias en la variable Y que está representada en ordenadas. No obstante existe una total simetría en las expresiones de la pendiente y ordenada en el origen de esta regresión de X sobre Y, pudiéndose obtener sus expresiones sin más que permutar x por y. Obteniéndose:

$$b' = \bar{x} - a'\bar{y} \qquad a' = \frac{S_{XY}}{S_Y^2}$$

Ejemplo: Si volvemos a los datos que hemos estado utilizando durante todo el capítulo, y pretendemos predecir los resultados del test de inteligencia a partir del rendimiento en la prueba, tendremos que calcular la recta de regresión de X sobre Y. Recordando los datos que hemos ido calculando, tenemos:

La media de X es 8 y su varianza vale 7,87

La media de Y es 3 y su varianza vale 2,27

La covarianza entre ambas variables es 3,67

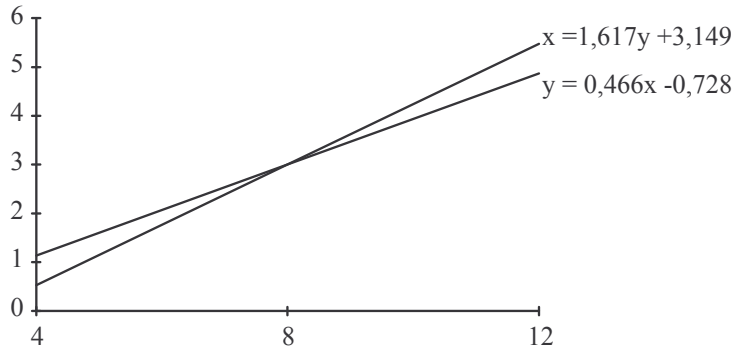
Por tanto la pendiente de la recta es 3,67/ 2,27 = 1,617

y la ordenada en el origen 8 - 1,617·3 = 3,149

Luego, la recta de regresión buscada es:

$$x = 1{,}617y + 3{,}149$$

Si representamos gráficamente ambas rectas apreciaremos la diferencia entre ambas:



Apéndice Matemático

Deducimos a continuación las expresiones de la pendiente y ordenada en el origen de la recta hace mínima la suma de los cuadrados de los residuos. Es decir buscamos los valores de a y b que hacen mínima la función:

$$H = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left[y_i - (ax_i + b) \right]^2$$

Para calcular el mínimo de esta función, determinaremos los puntos en que su derivada vale cero. Por consiguiente, derivando con respecto a b tenemos:

$$\frac{\partial H}{\partial b} = \sum_{i=1}^{n} 2 \cdot \left[y_i - (ax_i + b) \right] \cdot (-1)$$

igualando a cero:

$$\sum_{i=1}^{n} 2\left[y_i - (ax_i + b) \right](-1) = 0 \Rightarrow \sum_{i=1}^{n} \left[y_i - (ax_i + b) \right](-1) = 0$$

si lo expresamos como suma de sumatorios tendremos:

$$-\sum_{i=1}^{n} y_i + \sum_{i=1}^{n} ax_i + \sum_{i=1}^{n} b = 0 \Rightarrow \sum_{i=1}^{n} ax_i + \sum_{i=1}^{n} b = \sum_{i=1}^{n} y_i$$

sacando factor común, obtendremos la primera de las ecuaciones normales:

$$a\sum_{i=1}^{n} x_i + nb = \sum_{i=1}^{n} y_i$$

Derivando con respecto de a, tendremos:

$$\frac{\partial H}{\partial a} = \sum_{i=1}^{n} 2 \cdot \left[y_i - (ax_i + b) \right] \cdot (-x_i)$$

e igualando a cero:

$$\sum_{i=1}^{n} 2\left[y_i - (ax_i + b)\right](-x_i) = 0 \quad \Rightarrow \quad \sum_{i=1}^{n} \left[y_i - (ax_i + b)\right](-x_i) = 0$$

expresándolo como suma de sumatorios:

$$-\sum_{i=1}^{n} y_i x_i + \sum_{i=1}^{n} a x_i^2 + \sum_{i=1}^{n} b x_i = 0 \quad \Rightarrow \quad \sum_{i=1}^{n} a x_i^2 + \sum_{i=1}^{n} b x_i = \sum_{i=1}^{n} y_i x_i$$

sacando factor común, obtenemos la segunda de las ecuaciones normales:

$$a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i x_i$$

las ecuaciones normales constituyen un sistema de dos ecuaciones, del cual pueden obtenerse los valores de a y b. Despejando b de la primera de las ecuaciones normales, tenemos:

$$a\sum_{i=1}^{n} x_i + nb = \sum_{i=1}^{n} y_i \quad \Rightarrow \quad nb = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i \quad \Rightarrow \quad b = \frac{\sum_{i=1}^{n} y_i}{n} - a\frac{\sum_{i=1}^{n} x_i}{n}$$

con lo que obtenemos:

$$b = \bar{y} - a\bar{x}$$

Si en la segunda de las ecuaciones normales, dividimos ambos términos por n, tendremos:

$$a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i x_i \quad \Rightarrow \quad a\frac{\sum_{i=1}^{n} x_i^2}{n} + b\frac{\sum_{i=1}^{n} x_i}{n} = \frac{\sum_{i=1}^{n} x_i y_i}{n}$$

sustituyendo b por la expresión obtenida anteriormente, y teniendo en cuenta la expresión de la media, obtenemos:

$$a\frac{\sum_{i=1}^{n} x_i^2}{n} + (\bar{y} - a\bar{x})\bar{x} = \frac{\sum_{i=1}^{n} x_i y_i}{n} \quad \Rightarrow \quad a\frac{\sum_{i=1}^{n} x_i^2}{n} + \bar{y}\cdot\bar{x} - a\bar{x}^2 = \frac{\sum_{i=1}^{n} x_i y_i}{n}$$

sacando a factor común y reordenando los términos:

$$a\left(\frac{\sum_{i=1}^{n} x_i^2}{n} - \bar{x}^2\right) = \frac{\sum_{i=1}^{n} x_i y_i}{n} - \overline{xy} \quad \Rightarrow \quad a\cdot S_X^2 = S_{XY}$$

con lo que finalmente obtenemos:

$$a = \frac{S_{XY}}{S_X^2}$$